Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Optimizing *Shake-and-Bake* for proteins

Charles M. Weeks[a]* and Russ Miller[a,b]

[a]Hauptman–Woodward Medical Research Institute, Inc., 73 High Street, Buffalo, NY 14203, USA, and [b]State University of New York at Buffalo, Department of Computer Science and Engineering, Buffalo, NY 14260, USA

Correspondence e-mail: weeks@hwi.buffalo.edu

Received 20 April 1998
Accepted 30 September 1998

Shake-and-Bake is a direct-methods procedure which has provided *ab initio* solutions for protein structures containing as many as 1000 independent non-H atoms. This algorithm extends the range of conventional direct methods by repetitively, unconditionally and automatically alternating reciprocal-space phase refinement with filtering in real space to impose constraints. The application of *SnB* to protein-sized molecules is significantly affected by the choice made for certain critical parameters, including the number of peaks used for density modification, the choice of phase-refinement method and the number of refinement cycles. The effects of parameter variation have been studied for six protein structures, all of which are solvable by *Shake-and-Bake* using data at 1.1 Å or higher resolution. Solvability in the resolution range 1.2–1.4 Å appears to be enhanced by the presence of heavier atoms (S, Cl). Furthermore, it appears that in this range the ratio of refinement cycles and triplet phase invariants to atoms in the structure must be increased. Large structures lacking atoms of any element heavier than oxygen also require non-traditional parameter values.

1. Introduction

The potential for real-space constraints to improve phases in the context of small-molecule direct methods was recognized by Karle (1968), who found that even a relatively small chemically sensible fragment extracted by manual interpretation of an electron-density map could be parlayed into a complete solution by transformation back to reciprocal space followed by performing additional iterations of phase refinement. The tremendous increases in computer speed in recent years have made possible the development of a direct-methods multi-trial or 'multi-solution' technique (Germain & Woolfson, 1968) in which each trial structure is repeatedly cycled back-and-forth between real and reciprocal space, alternately performing optimization in each space. This computer-intensive process, which requires the use of two Fourier transforms during each cycle, is known as *Shake* (phase refinement) and *Bake* (density modification) (Weeks *et al.*, 1993; Miller *et al.*, 1993; Weeks, DeTitta *et al.*, 1994). This procedure has been described in detail in two recent reviews (Weeks & Miller, 1996, 1997). The ability to impose physically meaningful constraints in real space has increased the size of molecular structures amenable to phasing by direct methods from 100 to 1000 independent non-H atoms. The method known as iterative peak-list optimization (Sheldrick & Gould, 1995) has been patterned after *Shake-and-Bake* and relies even more heavily on real-space constraints.

Multi-trial direct-methods procedures require multiple sets of starting phases which can be subjected to a specified refinement protocol. In recent years, it has become routine to use a random-number generator to assign initial phase values (Baggio *et al.*, 1978; Yao, 1981). In the *Shake-and-Bake* procedure, phases are assigned initial values by first generating trial structures consisting of randomly positioned atoms (thereby imposing an atomicity constraint from the outset) and then computing structure factors. The tangent formula (Karle & Hauptman, 1956),

$$\tan(\varphi_{\mathbf{H}}) = \frac{\sum_{\mathbf{K}} |E_{\mathbf{K}} E_{\mathbf{H-K}}| \sin(\varphi_{\mathbf{K}} + \varphi_{\mathbf{H-K}})}{\sum_{\mathbf{K}} |E_{\mathbf{K}} E_{\mathbf{H-K}}| \cos(\varphi_{\mathbf{K}} + \varphi_{\mathbf{H-K}})}, \quad (1)$$

in either its original or a weighted form (Hull & Irwin, 1978), provides the means for phase refinement in conventional multi-solution phasing programs such as *MULTAN* (Main *et al.*, 1980) and *SHELXS* (Sheldrick, 1985).

On the other hand, *Shake-and-Bake* permits alternative optimization strategies during the phase-refinement step. A particularly good strategy is to use a parameter-shift search (Bhuiya & Stanley, 1963) to reduce the value of an objective function such as the minimal function

$$R(\varphi) = \Big(\sum_{\mathbf{H,K}} A_{\mathbf{HK}} \{\cos(\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{\mathbf{-H-K}}) - [I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}})]\}^2 \Big) \Big/ \sum_{\mathbf{H,K}} A_{\mathbf{HK}} \quad (2)$$

(Debaerdemaeker & Woolfson, 1983; Hauptman, 1991; DeTitta *et al.*, 1994). The minimal function expresses a relationship among phases related by triplet invariants that have the associated parameters (or weights)

$$A_{\mathbf{HK}} = (2/N^{1/2}) |E_{\mathbf{H}} E_{\mathbf{K}} E_{\mathbf{H+K}}|, \quad (3)$$

where the $|E|$s are the normalized structure-factor magnitudes and N is the number of atoms, assumed identical, in the unit cell. The minimal function $R(\varphi)$ is a measure of the mean-square difference between the values of the triplet invariants calculated using a trial set of phases and their expected values (given by the ratio of modified Bessel functions, I_1/I_0), as predicted by the conditional probability distribution of structure invariants (Cochran, 1955). It is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomorph. The relationship expressed by (2) is often written with negative quartets included. However, these invariants are not relevant to the study presented in this paper, since negative quartets are not expected to be reliable for protein-sized molecules. Experimentation has thus far confirmed that (i) when the minimal function is used actively in the phasing process and (ii) solutions actually occur, the final trial structure corresponding to the smallest value of $R(\varphi)$ is a solution. Therefore, $R(\varphi)$ is also an extremely useful figure of merit for selecting those trials which have converged to solution.

In the applications reported to date, automatic real-space electron-density map interpretation in the *Shake-and-Bake* procedure consists of selecting an appropriate number of the largest peaks (equal to or less than the expected number of atoms in the structure) to be used as an updated trial structure without regard to chemical constraints other than the powerful constraint of a minimum allowed distance. These peaks are then regarded as atoms and a structure-factor calculation imposes the atomicity constraint. If markedly unequal atoms are known to be present, appropriate numbers of peaks (atoms) can be weighted by the proper atomic numbers during transformation back to reciprocal space. Thus, *a priori* knowledge concerning the chemical composition of the crystal is utilized, but no knowledge of constitution is required or used during peak selection. It is useful to think of peak picking in this context as simply an extreme form of density modification appropriate when atomic resolution data are available. The entire dual-space refinement procedure is repeated for a predetermined number of cycles.

The *Shake-and-Bake* procedure has been implemented in an efficient and easy-to-use program entitled *SnB* (Miller *et al.*, 1994). Pertinent information concerning *SnB*, including the complete User's Manual, may be accessed from the home page on the WWW at www.hwi.buffalo.edu/SnB. The *SnB* user only needs to supply basic crystal data such as space group, cell constants and the contents of the asymmetric unit, as well as an input reflection file consisting of h, k, l and the normalized structure-factor magnitudes $|E|$. Cost-effective default values are provided for all the control parameters (displayed following each query) based on extensive experimentation with several known (mostly small-molecule) test structures (Weeks, DeTitta *et al.*, 1994; Weeks, Hauptman *et al.*, 1994; Chang *et al.*, 1997). The focus of this study is on (i) determining the optimum parameter values and operating conditions of the *SnB* program for small proteins having 300–600 independent non-H atoms and (ii) noting any changes in these parameters and conditions as the data resolution is lowered.

Table 1
Protein data sets used to test *SnB* parameters and operating conditions.

Protein	Space group	Protein atoms (n)	Total atoms	'Heavy' atoms	Maximum resolution range examined (Å)
Vancomycin	$P4_32_12$	202	258	Cl8	0.9–1.4
Conotoxin EpI	$I4$	248	289	S10	1.1–1.4
Gramicidin A	$P2_12_12_1$	272	317	—	0.86–1.1
Crambin	$P2_1$	327	~400	S6	0.83–1.2
Rubredoxin	$P2_1$	395	497	FeS6	1.0–1.1
Scorpion toxin II	$P2_12_12_1$	508	624	S8	0.96–1.2

2. Materials and methods

The effects of changes in parameter values or operating conditions were studied by examining the frequency of solutions and the rate of convergence for sets of 1000 or more trial structures for the six proteins listed in Table 1. For purposes of this communication, a polypeptide containing more than 200 non-H atoms was considered to be a 'protein'. These structures crystallize in a variety of space groups, and the total

numbers of independent non-H atoms in the protein alone and in the protein plus solvent and ligand molecules are indicated, as well as the presence of S, Cl or Fe atoms. The data for five of these test structures were measured to a resolution of 1.0 Å or higher, and these data were truncated to various resolutions between 1.1 and 1.5 Å in order to study the effects of lowering the resolution. The resolution range defines the resolution of the full data set and the lowest resolution of the corresponding truncated data set for which *SnB* solutions could be obtained by the procedures described here. Truncated data are of higher quality than data which do not exceed a particular resolution. Therefore, results obtained with truncated data should be regarded as best-case scenarios.

The glycopeptide antibiotic vancomycin was originally solved simultaneously and independently in two laboratories (Loll *et al.*, 1997; Schafer *et al.*, 1996) using different programs that implement the *Shake-and-Bake* paradigm, namely *SnB* (Miller *et al.*, 1994) and Sheldrick and Gould's variant of the *Shake-and-Bake* procedure which employs iterative peak-list optimization (Sheldrick & Gould, 1995). The 289-atom structure of conotoxin EpI was also solved originally by *SnB* (Hu *et al.*, 1998). The 64-residue scorpion toxin (Tox II), one of the largest successful *Shake-and-Bake* applications so far, had been solved previously by conventional isomorphous replacement techniques, but it was redetermined using *SnB* (Smith *et al.*, 1997) without knowledge of the exact number of residues or the amino-acid sequence. The remaining test structures [crambin (Hendrickson & Teeter, 1981), gramicidin A (Langs, 1988) and rubredoxin (Dauter *et al.*, 1992)] were all determined originally by other methods. However, these three structures were redetermined using *SnB* in order to demonstrate the feasibility of obtaining automated *ab initio* direct-method solutions for small proteins. The application to the crambin data has been described in detail elsewhere (Weeks *et al.*, 1995).

All experiments were conducted on a network of SGI R10000 Indigo 2 workstations using a prototype of *SnB* v2.0 (Weeks & Miller, 1999). In addition to new scientific options, *SnB* v2.0 has improved crystallographic core routines which have been redesigned and written in a more tightly integrated form which serves to improve the efficiency of the procedure. The parameters that were evaluated included the number of refinement cycles, phases, triplet structure invariants (or ratio of invariants to phases) and peaks selected during real-space filtering. When comparing results for different structures, it is helpful to express all of these quantities as a function of n, the number of independent non-H atoms in the asymmetric unit. For purposes of this study, n was taken to be the number of such atoms in the protein molecule itself. Unless stated otherwise, trial structures were refined for n cycles using $10n$ phases and $100n$ triplet invariants. Initial trial structures containing 100 atoms were used to generate the starting phases for the first cycle. When 'heavier' atoms (S, Cl, Fe) were present, a corresponding number of the largest peaks were weighted in the structure-factor calculations by the atomic numbers of these elements, and the remaining peaks were treated as nitrogen.

Both the parameter-shift and tangent-formula phase-refinement procedures were applied to the six protein data sets. The phases are ranked in decreasing order with respect to the values of their associated normalized structure-factor magnitudes ($|E|$). Parameter-shift refinement proceeds by allowing each phase, in turn, to take on two or more alternative values. The value of $R(\varphi)$ (the minimal function) is determined for each such phase. The phase value yielding the minimum $R(\varphi)$ is taken as the refined value and immediately replaces the old value. Based on experimentation with known small-molecule structures, it has been determined that a good parameter-shift strategy is to perform a maximum of two 90° phase shifts and to make three complete iterations or passes through the phase set during each *Shake-and-Bake* cycle (Weeks, DeTitta *et al.*, 1994). The notation PS(90°,2,3) is an abbreviation denoting these operating conditions. When parameter-shift phase refinement is applied in centrosymmetric space groups, only a single shift of 180° is required for each phase. In the acentric space groups of interest in this study, phases having restricted values may either be constrained to these values (restricted PS) or treated as general phases (unrestricted PS or standard PS). In either case, space-group restrictions are reapplied during the structure-factor calculation. In the case of tangent refinement, the minimal function was also computed, but was used only as a figure of merit. The tangent formula was implemented in its original form (Karle & Hauptman, 1956). As in the parameter-shift case, there was feedback of each refined phase value which could be used immediately in the refinement of the next phase rather than waiting until the beginning of the next cycle. Previous studies (Chang *et al.*, 1997) indicated that it was best to perform only one or two iterations of tangent refinement per cycle for structures containing more than 100 independent atoms.

In the following discussion, different *Shake-and-Bake* protocols will be compared on the basis of (i) success rate and (ii) cost-effectiveness (*i.e.* efficiency). A completed trial structure is considered to be a solution if it has a close match between peak positions and the true atomic positions for some choice of origin and enantiomorph, as well as a mean phase error of 40° or less. The term 'success rate' is used to refer to the percentage of trial structures that become solutions over the course of refinement. In space groups such as $P2_12_12_1$ and $P4_32_12$, where there are only a few possible discrete origin positions, completed *Shake-and-Bake* trials for known structures can be rapidly screened for solutions by examining the mean phase error or average absolute value of the deviations of the phases from their known values calculated using final refined coordinates and thermal parameters. In space groups $P2_1$ and $I4$, where the origin can be moved along one axis, mean phase errors were computed at intervals of one-hundredth of the cell length along that axis. Since the minimal function $R(\varphi)$ is used to identify probable solutions for unknown structures, consideration of both the mean phase error and $R(\varphi)$ permits trials to be categorized as true, false or missed solutions or as non-solutions.

Although the measurement of success rates at the end of a fixed number of cycles provides an important indication as to the effectiveness of a particular method, success rate by itself provides an incomplete comparison of two refinement protocols because it does not take into account the computational effort (running time) needed to produce the solutions. The relative efficiency of two procedures can be compared on the basis of the cost-effectiveness (*i.e.* the number of solutions obtained per unit of time using a particular computing platform). The cost-effectiveness achieved in a particular run depends on the number of cycles actually performed. In this study, a large number of cycles was used so that the cycle with maximum cost-effectiveness could be ascertained. The cost-effectiveness for the default number of cycles, or its maximum value along with an indication of the cycle at which this maximum occurred, provides meaningful comparisons of different experimental conditions.

3. Results

3.1. Atom:phase:invariant ratio

Extensive study of several small-molecule data sets suggested that an atom:phase:triplet invariant ratio of 1:10:100 is appropriate for the *Shake-and-Bake* procedure with parameter-shift phase refinement (Weeks, DeTitta *et al.*, 1994). This ratio was confirmed by the behavior of the crambin and gramicidin A data sets, and these results are reported in Table 2. Although the success rate may be higher when more phases are used, the refinement process has peak efficiency when the atom:phase ratio is near 1:10. Considerable latitude seems to be acceptable with atom:phase ratios in the range 1:7 to 1:15 yielding similar results. On the other hand, both the success and the efficiency decrease significantly when too many invariants are used (phase:invariant ratio > 1:20).

Table 2
Effects of the atom:phase:invariant ratio on the success rate and cost-effectiveness (CE) of 2000 *Shake-and-Bake* trials for crambin (300 cycles, 100 peaks) and gramicidin A (500 cycles, 200 peaks).

(*a*) Phase:invariant ratio (*P*:*I*) fixed at 1:10.

		Crambin		Gramicidin A	
Phases	Invariants	Success rate (%)	Maximum CE (solutions h^{-1})	Success rate (%)	Maximum CE (solutions h^{-1})
1000	10000	1.1	0.4	0.5	0.04
2000	20000	3.9	1.8	2.0	0.15
3000	30000	4.4	2.1	3.0	0.18
4000	40000	4.6	1.8	3.3	0.16
5000	50000	4.6	1.5	3.5	0.15

(*b*) Number of phases fixed at 3000.

		Crambin		Gramicidin A	
Phases	Invariants	Success rate (%)	Maximum CE (solutions h^{-1})	Success rate (%)	Maximum CE (solutions h^{-1})
3000	15000	4.2	1.6	1.3	0.09
3000	30000	4.4	2.1	3.0	0.18
3000	45000	4.2	1.8	2.9	0.18
3000	60000	3.5	1.3	3.1	0.17
3000	90000	3.1	1.3	1.1	0.05
3000	150000	2.9	0.6	0.7	0.02
3000	300000	2.2	0.1	0.2	0.01
3000	600000	—	—	0.0	0.00

3.2. Choosing the number of peaks

Previous experience has shown that it is sufficient to use 100 randomly positioned atoms as an initial trial structure for the first cycle. During subsequent cycles, choosing n peaks to recycle through the procedure gives optimum success rates for smaller structures. However, for large structures containing a significant number of atoms with low occupancy or high thermal motion, recycling structures containing fewer than n peaks might be expected to give a better performance. This hypothesis was tested for the six test structures, and the results are reported in Table 3. In all but one case, the optimum number of peaks to use appears to be significantly less than half the number of atoms in the protein molecule itself. The exception to this rule is gramicidin A, the only one of the six structures containing solely first-row elements.

The variation in success rate, as a function of the number of peaks selected, is especially significant for scorpion toxin II (Tox II). The original *SnB* study of this structure (Smith *et al.*, 1997) consisted of processing trial structures for 255 (or $n/2$) cycles of unrestricted parameter shift (maximum of two 90° phase shifts; three iterations) with 400 peaks selected for use in the structure-factor calculation. This yielded one solution from 1619 trials. In one recent experiment, approximately 18000 such trials were generated, and the lowest phase errors obtained after 255 cycles were 58, 74 and 79°, one of which was detectable based on the minimal function (*i.e.* it had the lowest value). These three trials were then subjected to further refinement and found to converge to solution (mean phase error less than 40° after 270, 295 and 320 cycles, respectively). In a second recent experiment, approximately 3300 400-peak trials were refined for 500 cycles. Three of these trials converged to solution, but only after more than 480 *SnB* cycles. Therefore, one may conclude that the original *SnB* solution of Tox II was indeed serendipitous! In contrast, the 200-peak trials are much more likely to lead to solution, emphasizing the importance of using fewer peaks in this case.

3.3. Weighting of peaks

When atoms with atomic numbers greater than 10 are present, the *SnB* user has the option of weighting the appropriate number of largest peaks in the structure-factor calculations more heavily. Unequal weighting has resulted in accelerated convergence to solution for small molecules having a few Cl atoms (Weeks, De Titta *et al.*, 1994). Therefore, the default option is to weight the appropriate number of the largest peaks with the atomic numbers of elements heavier than oxygen. On account of the strong influence which such atoms appear to exert on the progress of refinement, experiments designed to address the following questions were conducted with the crambin data (crambin contains six S atoms). Is it more efficient to try to determine the sulfur

Table 3
Influence of the number of peaks used in the structure-factor calculation on the success rate of *Shake-and-Bake* with unrestricted parameter-shift (90°,2,3) phase refinement and a *P*:*I* ratio of 1:10.

The minimum $|E|$ values for the phases used and the minimum A values for the triplet invariants used are indicated.

						Success rate (%)							
						Peaks used							
Protein	Trials	Cycles (∼n)	Phases	E min	A min	25	50	100	150	200	250	300	400
Vancomycin	1000	200	2000	1.50	0.58	0.8	0.4	0.6	0.3	0.2	0.2	—	—
Conotoxin EpI	1000	250	1900	1.25	0.46	44.0	53.0	52.0	50.0	45.0	42.0	—	—
Gramicidin A	5000	275	3000	1.40	0.50	—	0	0.3	0.5	1.1	0.9	0.7	—
Crambin	2000	300	3000	1.47	0.60	3.0	4.3	4.8	3.9	3.3	2.6	3.4	—
Rubredoxin	1000	400	4000	1.23	0.51	5.0	5.7	6.2	6.0	5.4	5.1	3.9	3.4
Toxin II	>2000	500	5000	1.34	0.39	—	—	1.0	—	1.4	—	0.4	0.1

Table 4
Effects of selecting fewer peaks or treating fewer peaks as sulfur in the early *SnB* cycles for crambin.

In each experiment, a total of 150 parameter-shift cycles were performed for 3000 phases using 30 000 triplet invariants.

	Peak-selection conditions						
Experiment	Cycles	Peaks	Sulfurs	Trials	Number of solutions	Success rate (%)	Max CE (solutions h^{-1})
Control	1–150	100	6	11600	407	3.5	1.5
1	1–150	100	0	200	22	1.1	0.4
2	1–10	100	0	2000	70	3.5	1.2
	11–150	100	6				
3	1–10	100	2	2000	58	2.9	1.5
	11–20	100	4				
	21–150	100	6				
4	1–25	25	6	4000	123	3.1	1.3
	26–150	100	6				
5	1–50	25	6	4000	127	3.2	1.1
	51–150	100	6				
6	1–5	25	6	4000	129	3.2	1.5
	6–10	50	6				
	11–150	100	6				
7	1–5	25	2	3875	124	3.2	1.6
	6–10	50	4				
	11–150	100	6				

positions first by choosing fewer peaks in the early cycles? Is it more efficient to treat fewer peaks as sulfurs in the early cycles before the heavier atom positions are well established? Based on the data reported in Table 4, the answers to both questions are negative, at least for this structure. None of the experiments produced a higher success rate or more solutions per hour than did the control run in which 100 peaks were selected in each cycle and the largest six peaks were always weighted as sulfurs in the structure-factor calculation. The experiment in which the knowledge that sulfur is present was disregarded and all peaks were weighted equally produced the worst results. This confirms the observation, made for small-molecule data, that it is preferable to utilize such information. The suggestion has been made (G. Sheldrick, personal communication) that, in working with structures that are expected to have several disulfide bonds, it may be efficient to test the largest peaks for appropriate distances in an early cycle and terminate those trials lacking the required number of such distances. The data reported in Table 5 show that, in this case, none of the early termination conditions that were tested was more efficient than the control conditions.

3.4. Phase refinement

In Table 6, both restricted and unrestricted parameter-shift phase refinement (defined in §2) are compared to phase refinement consisting of one or two iterations of tangent refinement. It can be seen that for all six proteins, unrestricted parameter shift is superior to restricted parameter shift and both forms of parameter shift generally produce more solutions than the tangent formula. Again, the results for gramicidin A stand out. No tangent-based solutions were obtained after 275 cycles, in contrast to the 17 solutions obtained with unrestricted parameter shift. In fact, all 2000 gramicidin A trials in this experiment were subjected to a total of 500 refinement cycles, and only one tangent solution was obtained (at about cycle 350). After 500 cycles, there were 53 unrestricted parameter-shift solutions. As shown by the results presented in Table 7, gramicidin A continues to behave differently when the question of the optimum number of refinement cycles is considered. It is most efficient to perform $n/2$ or fewer cycles for four of these six proteins. The exceptions are gramicidin and Tox II, the largest structure, for which $3n/4$ cycles is at least as good as $n/2$ cycles. Therefore, it may be advisable to perform at least $3n/4$ cycles for larger structures so as to ensure that solutions are not missed altogether. The efficiency of solution for gramicidin A is highest at 500 cycles (∼$1.75n$), the maximum tested. The suggestion is that large structures

Table 5
Effects on efficiency (CE) of early trial termination based on the absence of distances in the range 1.7–2.4 Å (possible disulfide bonds) between the highest 12 or 25 peaks.

Distances were checked after every 10, 20, 30 or 50 cycles, as indicated. The 11 600 crambin trials were refined for 150 parameter-shift cycles using 3000 reflections and 30 000 invariants, 100 peaks were selected in each cycle and the largest six peaks were treated as sulfur in the structure-factor calculations. The control run with no distance check produced solutions at the rate of 1.5 h^{-1}.

		Maximum CE (solutions h^{-1})	
Checkpoint cycles	Minimum required S—S bonds	12 'S' peaks	25 'S' peaks
Every 10	1	0.6	1.3
	2	0.1	1.0
	3	0.0	0.2
Every 20	1	1.4	1.4
	2	0.8	1.3
	3	0.6	0.8
Every 30	1	1.5	1.4
	2	1.3	1.5
	3	0.9	1.3
Every 40	1	1.6	1.4
	2	1.4	1.4
	3	1.2	1.3

lacking any atoms heavier than oxygen are especially likely to require many cycles for convergence to solution to be achieved. This also appears to be true for the smaller, but difficult, structure ternatin ($n \simeq 100$) for which oxygen is also the heaviest element (Miller *et al.*, 1993).

3.5. Effect of data resolution

The results obtained when the protein test-data sets were truncated to resolutions in the range 1.1–1.4 Å are summarized in Table 8. All of these structures are solvable at 1.1 Å, but success rates decrease as resolution decreases. However, success rates may be increased, or the range of resolutions over which solutions are obtainable may be extended, by making suitable adjustments to the values of certain control parameters. In particular, increasing the ratio of triplet invariants to phases or increasing the number of *SnB* cycles are especially beneficial, as the resolution decreases to 1.2 Å or lower. Fig. 1(*a*) illustrates how either of these parameter adjustments increases the success rate for vancomycin at 1.1 Å, and Fig. 1(*b*) shows the number of solutions occurring in successive blocks of $n/2$ cycles for three of the other structures. It is interesting to note that, using 500 SnB cycles with truncated 1.1 Å toxin II data and a phase:invariant ratio (*P*:*I*) of 1:10, no solutions were found for 2000 trials. However, when the ratio was increased to 1:20, the success rate was 1 in 200 trials. Similar results were obtained at 1.1 Å using the crambin data for which the 300-cycle success rates were 0.2 and 1.1% for *P*:*I* ratios of 1:10 and 1:30, respectively. Using the full (0.86 Å) gramicidin A data set, the 275-cycle success rate (*P*:*I* ratio = 1:10) was 1.1%, but the earliest solution (2000 trials) occurs at cycle 327 using the 1.1 Å truncated data and a *P*:*I* ratio of 1:25.

The information presented in Table 9 for vancomycin shows that, in agreement with the results obtained for the full 0.9 Å data set, parameter shift is superior to the tangent formula as a phase-refinement method for the 1.1 Å truncated data. At present, the available evidence concerning the relationship (if any) between resolution and the number of peaks selected during the density-modification step is conflicting. As shown above for the full data sets, the optimum number of peaks recycled as atoms for crambin and toxin II are 100 and 200,

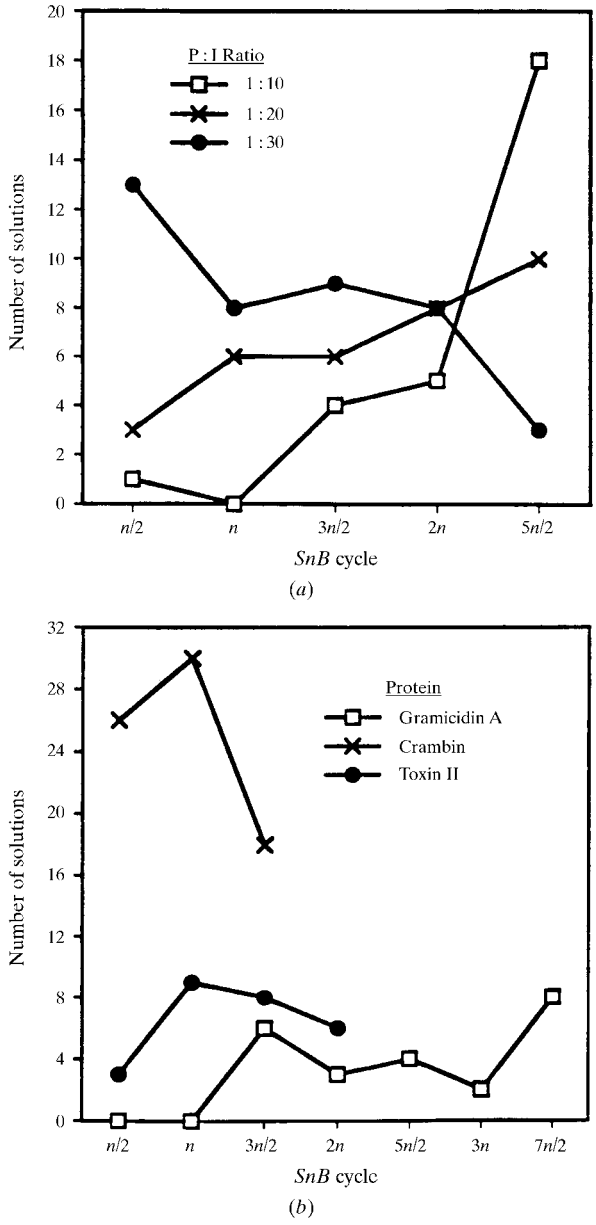

Figure 1
Numbers of unrestricted PS solutions occurring during blocks of $n/2$ cycles for data truncated to 1.1 Å resolution. (*a*) Vancomycin, 2500 trials, 2000 phases and 100 peaks. (*b*) Gramicidin A, 2000 trials, 3000 phases, *P*:*I* ratio 1:25 and 200 peaks; crambin, 5000 trials, 3000 phases, *P*:*I* ratio 1:30 and 100 peaks; scorpion toxin II, 2000 trials, 5000 phases, *P*:*I* ratio 1:20 and 200 peaks.

Table 6
Success rates for different phase-refinement options using a *P:I* ratio of 1:10.

The number of peaks selected for each structure was chosen based on the optimum values reported in Table 3.

					Success rates			
					Parameter-shift (90°,2,3)			
Protein	Trials	Cycles ($\simeq n$)	Phases	Peaks	Unrestricted (%)	Restricted (%)	Tangent formula (1) (%)	Tangent formula (2) (%)
Vancomycin	2500	200	2000	100	0.6	0.4	0.3	0.4
Conotoxin EpI	1000	250	1900	50	53.0	44.0	25.0	31.0
Gramicidin A	2000	275	3000	200	1.1	0.5	0	0
Crambin	2000	300	3000	100	4.8	3.7	2.2	0.6
Rubredoxin	1000	400	4000	150	6.0	5.2	4.0	0.9
Toxin II	>2000	500	5000	200	1.4	1.0	0.7	0

Table 7
Success rates as a function of the number of unrestricted parameter-shift (90°,2,3) refinement cycles.

The numbers of cycles are expressed in terms of *n*, the number of independent non-H atoms in the protein. The numbers of trials, phases, invariants and Fourier peaks used for each structure are given in Table 6.

	Success rate with increasing *SnB* refinement cycles (%)					
Protein	0.25*n*	0.5*n*	0.75*n*	*n*	1.25*n*	1.5*n*
Vancomycin	0.1	0.4†	0.4	0.6	0.7	—
Conotoxin EpI	27.0†	40.0	48.0	53.0	—	—
Gramicidin A	0	0.4	0.6	0.9	1.2	2.0†
Crambin	3.1†	4.1	4.6	4.8	—	—
Rubredoxin	4.6†	5.5	5.9	6.0	—	—
Toxin II	0.05	0.5	1.0†	1.4	—	—

† Most efficient (cost-effective) number of cycles.

Table 8
Success rates for truncated data using unrestricted PS (90°,2,3) phase refinement.

Resolution (Å)	Protein	Trials	Cycles	Phases	Peaks	*P:I* ratio	Success rate (%)
1.1	Vancomycin	2500	*n*	2000	100	1:10	0.04
						1:20	0.4
						1:30	0.8
	Gramicidin A	2000	*n*	3000	200	1:25	0
			2*n*			1:25	0.4
	Crambin	5000	*n*	3000	100	1:10	0.2
					100	1:30	1.1
					200	1:30	0
	Rubredoxin	1000	*n*	2000	50	1:10	5.0
	Toxin II	2000	*n*	5000	200	1:10	0
						1:20	0.5
1.2	Vancomycin	2500	*n*	2000	100	1:60	0.5
	Conotoxin EpI	4000	*n*	1440	50	1:10	38.0
	Crambin	4354	1.5*n*	2000	100	1:50	0.02
	Toxin II	753	2*n*	3000	100	1:80	0.7
		2557			200	1:80	0
1.3	Vancomycin	2500	*n*	1500	100	1:80	1.0
	Conotoxin EpI	5000	*n*	1125	50	1:10	14.7
1.4	Vancomycin	1914	*n*	1500	100	1:80	1.4
	Conotoxin EpI	1000	*n*	1250	50	1:50	3.6

respectively. Crambin results at 1.1 Å (1.1% success when 100 peaks were selected *versus* zero solutions out of 5000 trials when 200 peaks were selected) were consistent with the observations for the full data. On the other hand, there are zero solutions out of 2500 toxin II trials when 200 peaks are selected at 1.2 Å, but there are five solutions out of 750 trials when only 100 peaks were selected (1000 cycles; *P:I* ratio 1:80).

It is interesting to note that the two test structures for which solutions have not been achieved at a resolution lower than 1.1 Å are gramicidin A, which has no atoms heavier than oxygen, and rubredoxin, which has a cluster of 'heavier' atoms. In contrast, the two structures (vancomycin, conotoxin EpI) with the most well separated heavier atoms (Cl, S) can be solved with truncated 1.3 or 1.4 Å data. It seems likely that the presence of several heavier atoms facilitates solvability at lower-than-normal resolution. At 1.4 Å, there is some overlap of minimal function values for solutions and non-solutions such that the histogram does not have its typical bimodal shape, but the best values still earmark solutions. At 1.5 Å, there are no solutions for any of the test structures. Although some 1.5 Å trials for vancomycin and conotoxin EpI do have mean phase errors in the 50–60° range, these trials are not identifiable on the basis of the minimal function.

4. Conclusions

The ultimate potential of the *Shake-and-Bake* approach to the *ab initio* structure determination of macromolecules is unknown. Based on the study of six proteins reported here, it is possible, however, to suggest guidelines that maximize the probability of obtaining a solution for a structure containing several hundred atoms. These recommendations are summarized in Table 10 for 1 Å (or higher resolution) protein data, and they differ somewhat from the default values used in *SnB* v1.5 (Weeks, Hauptman *et al.*, 1994; Chang *et al.*, 1997). If several atoms of moderately 'heavy' elements (S, Cl, Fe) are present, it may be best to select a

Table 9
Success rates of different phase-refinement methods for 2500 vancomycin trials using 1.1 Å truncated data, 2000 phases, a *P*:*I* ratio of 1:20 and selecting 100 peaks in each cycle.

	PS (90°,2,3)		
Cycles	Unrestricted (%)	Restricted (%)	Tangent formula (1) (%)
n	0.4	0.5	0.2
2*n*	0.9	0.8	0.3

Table 10
Recommendations for *SnB* parameters for a structure with *n* non-H protein atoms.

Resolution	Parameter	Recommendation
1.0 Å or higher	Phases	10*n*
	Triplet invariants	100*n*
	Peaks	0.4*n* if several 'heavy' atoms present 0.8*n* if all C, N, O
	Cycles	*n*/2 if $n < 400$ and 'heavy' atoms present *n* otherwise
	Phase refinement	Unrestricted parameter shift PS (90°,2,1) if *P*1 PS (90°,2,3) otherwise
1.1–1.2 Å	Phases	<10*n* (keep minimum $E > 1.2$)
	Triplet invariants	200*n*–500*n*
	Cycles	*n*–2*n*

number of peaks equivalent to about 40% of the non-H protein atoms in the asymmetric unit. If no such elements are present, the 80% rule previously suggested is a better choice. Undoubtedly neither of these two extremes is appropriate in all cases. There was an absence of appropriate test data sets having fewer than six 'heavy' atoms, so it was not possible to determine an optimum number of peaks for such cases. The recommendations in Table 10 provide a starting point for consideration, but it is probably prudent to 'interpolate' based on the number of heavy atoms present. It appears that it is always best to use parameter shift as a phase-refinement method rather than the tangent formula, especially if only C, N and O atoms are present. Furthermore, the parameter-shift refinement should always be unrestricted, regardless of space-group phase restrictions. If the structure is larger than 500 non-H atoms or contains only the lighter elements, it appears to be prudent to perform at least as many refinement cycles as there are independent non-H protein atoms (*i.e.* *n* cycles).

If the resolution of the data is 1.1 Å or less, it is wise to increase further the number of refinement cycles and/or increase the phase:invariant ratio to 1:20 (at least). Furthermore, since it is generally unwise to use reflections with lower *E* values (*e.g.* <1.2) in direct-methods phasing, at some point in the 1.1–1.2 Å range it is typically necessary to reduce the number of reflections being phased to less than 10*n*. Of course, there will always be structures which appear to violate these guidelines. For example, conotoxin EpI solves readily with only 100*n* triplet invariants and *n* refinement cycles, even at 1.3 Å resolution. However, if good computing facilities are available, it is better to err on the conservative side rather than risk a 0% success rate (*e.g.* 1.1 Å gramicidin A with only *n* cycles or scorpion toxin II with only 100*n* invariants). In general, applications to 1.3 or 1.4 Å data are not recommended unless there are a significant number of atoms heavier than oxygen (*e.g.* vancomycin, conotoxin EpI).

In Table 11, the throughput for the six proteins in terms of trials per day is presented for *SnB* v2.0. Assuming that *n*/2 cycles are performed and an optimum number of peaks selected in each cycle, it can be determined from the expected success rates and trial turnover that there is a high probability that a solution will be found in a single day for all of these structures, with the exception of scorpion toxin II, using a single SGI R10000 processor. If more cycles ($\sim n$) are performed for each trial processed, there is a strong likelihood that even Tox II would be solved on a single workstation in a single day. Thus, given the current level of computing technology, *Shake-and-Bake* is a practical method for solving structures of this size and complexity provided that adequate data have been measured.

Table 11
SnB v2.0 success rate and throughput for the full data sets (given as trials per day) using unrestricted parameter-shift (90°,2,3) phase refinement, a *P*:*I* ratio of 1:10, the optimum number of peaks (highest success rate) as given in Table 3 and *n*/2 cycles.

The numbers of trials per day are for a single SGI R10000 processor.

Structure	Phases	Cycles (*n*/2)	Peaks	Success rate (%)	Trials per day	Solutions per day
Vancomycin	2000	100	100	0.4	391	1.6
Conotoxin EpI	1900	125	50	40.0	274	110
Gramicidin A	3000	135	200	0.4	572	2.3
Crambin	3000	150	100	4.1	1029	42
Rubredoxin	4000	200	150	5.5	294	16
Toxin II	5000	250	200	0.5	109	0.5

The *Shake-and-Bake* algorithm and the *SnB* program have been made possible by the financial support of grants GM-46733 from NIH and IRI-9412415 from NSF. The authors would like to express their appreciation to Steve Gallo for all his work in the development of *SnB* v1.5, to Jan Pevzner for assisting with the coding of the inverse Fourier transform in *SnB* v2.0, to Brett Miller, Hongliang (Jimmy) Xu and Li Li for assistance in obtaining and tabulating some of the results and to Professor Herbert A. Hauptman for his continued inspiration and support.

References

Baggio, R., Woolfson, M. M., Declercq, J.-P. & Germain, G. (1978). *Acta Cryst.* A**34**, 883–892.
Bhuiya, A. K. & Stanley, E. (1963). *Acta Cryst.* **16**, 981–984.
Chang, C.-S., Weeks, C. M., Miller, R. & Hauptman, H. A. (1997). *Acta Cryst.* A**53**, 436–444.
Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.
Dauter, Z., Sieker, L. C. & Wilson, K. S. (1992). *Acta Cryst.* B**48**, 42–59.

Debaerdemaeker, T. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 193–196.
DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Germain, G. & Woolfson, M. M. (1968). *Acta Cryst.* B**24**, 91–96.
Hauptman, H. A. (1991). *Crystallographic Computing 5: From Chemistry to Biology*, edited by D. Moras, A. D. Podnarny & J. C. Thierry, pp. 324–332. Chester/Oxford: IUCr/Oxford University Press.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Hu, S.-H., Loughnan, M., Miller, R., Weeks, C. M., Blessing, R. H., Alewood, P. F., Lewis, R. J. & Martin, J. L. (1998). *Biochemistry*, **37**, 11425–11433.
Hull, S. E. & Irwin, M. J. (1978). *Acta Cryst.* A**34**, 863–870.
Karle, J. (1968). *Acta Cryst.* B**24**, 182–186.
Karle, J. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635–651.
Langs, D. A. (1988). *Science*, **241**, 188–191.
Loll, P. J., Bevivino, A. E., Korty, B. D. & Axelsen, P. H. (1997). *J. Am. Chem. Soc.* **119**, 1516–1522.
Main, P., Fiske, S. J., Hull, S. E., Lessinger, L., Germain, G., Declercq, J. P. & Woolfson, M. M. (1980). *MULTAN80. A System of Computer Programs for the Automatic Solution of Crystal Structures from X-ray Diffraction Data.* Universities of York, England, and Louvain, Belgium.
Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.
Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.
Schafer, M., Schneider, T. R. & Sheldrick, G. M. (1996). *Structure*, **4**, 1509–1515.
Sheldrick, G. M. (1985). *SHELXS86. Program for the Solution of Crystal Structures.* University of Göttingen, Germany.
Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.
Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A. & Miller, R. (1997). *Acta Cryst.* D**53**, 551–557.
Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Weeks, C. M., DeTitta, G. T., Miller, R. & Hauptman, H. A. (1993). *Acta Cryst.* D**49**, 179–181.
Weeks, C. M., Hauptman, H. A., Chang, C.-S. & Miller, R. (1994). *Am. Crystallogr. Assoc. Trans.* **30**, 153–161.
Weeks, C. M., Hauptman, H. A., Smith, G. D., Blessing, R. H., Teeter, M. M. & Miller, R. (1995). *Acta Cryst.* D**51**, 33–38.
Weeks, C. M. & Miller, R. (1996). In *Crystallographic Computing 7: Proceedings of the Macromolecular Crystallography Computing School*, edited by P. Bourne & K. Watenpaugh. Bellingham, WA: Western Washington University. Also available at http://www.sdsc.edu/Xtal/IUCr/CC/School96/.
Weeks, C. M. & Miller, R. (1997). *Proceedings of the CCP4 Study Weekend*, edited by K. S. Wilson, G. Davies, A. W. Ashton & S. Bailey, pp. 139–146. Warrington: Daresbury Laboratory. Also available at http://www.dl.ac.uk/CCP/CCP4/main.html.
Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**. In the press.
Yao, J.-X. (1981). *Acta Cryst.* A**37**, 642–664.